CSC321 Programming Assignment 1

Vitaly Topekha

February 7, 2018

1 Part 1: Network architecture

1. word_embedding weights = $250 * 16 = 4000$
   embed_to_hid weights = $128 * 3 * 16 = 6144$
   hid_bias =128
   hid_to_output weights = $250 * 128 = 32000$
   output_bias = 250
   Total trainable parameters = $4000 + 6144 + 128 + 32000 + 250 = 42,522$
   We can observe that "hid to output" weights have the most trainable parameters.

2. 250^4 = 3,906,250,000

2 Part 2: Training the model

```
loss_derivative[2, 5] 0.0013789153741
loss_derivative[2, 121] -0.999459885968
loss_derivative[5, 33] 0.000391942483563
loss_derivative[5, 31] -0.708749715825

param_gradient.word_embedding_weights[27, 2] -0.298510438589
param_gradient.word_embedding_weights[43, 3] -1.13004162742
param_gradient.word_embedding_weights[22, 4] -0.211118814492
param_gradient.word_embedding_weights[2, 5] 0.0

param_gradient.embed_to_hid_weights[10, 2] -0.0128399532941
param_gradient.embed_to_hid_weights[15, 3] 0.0937808780803
param_gradient.embed_to_hid_weights[30, 9] -0.16837240452
param_gradient.embed_to_hid_weights[35, 21] 0.0619595914046

param_gradient.hid_bias[10] -0.125907091215
param_gradient.hid_bias[20] -0.389817847348

param_gradient.output_bias[0] -2.23233392034
```

```
param_gradient.output_bias[1] 0.0333102255428
param_gradient.output_bias[2] -0.743090094025
param_gradient.output_bias[3] 0.162372657748
```

3 Part 3: Analysis

1. ```
   >>> lm.find_occurrences('you', 'make', 'me')
   The tri-gram "you make me" was followed by the following words in the training set:
   some (1 time)
   >>> model.predict_next_word('you', 'make', 'me')
   you make me ? Prob: 0.17252
   you make me . Prob: 0.12092
   you make me do Prob: 0.05404
   you make me see Prob: 0.05091
   you make me up Prob: 0.02895
   you make me , Prob: 0.02700
   you make me play Prob: 0.02631
   you make me like Prob: 0.02626
   you make me in Prob: 0.02572
   you make me think Prob: 0.02218
   ```

2. Most of the words in each cluster can be used in the same context (e.g.: five, four, three, several), many of them are synonyms (e.g.: only, just) and most are represented by the same parts of speech (e.g.: they, we, you, i, she, he).

3. The words 'new' and 'york' are located pretty far apart from each other on the graph. I believe that is the case, because even though they are used together in a single context of "new york city", most of the times we would observe that word 'new' would be closer to 'old' in most case in terms of context. However that also depends on the data set. if you take data from "New York Times" that is a no-brainer that they will be closer.

4. Words 'government' and 'university' seem to be located closer to each other on the graph than words 'government' and 'political'. I think the reason why it happened is that 'government' and 'university' are both nouns. Even though that it makes more sense that 'political' and 'government' should be closer to each other in terms of context, they are different parts of speach and have different function in a sentence.
   ```